STAR MDC1 Experience and Revised Computing Requirements

STAR COMPUTING

Torre Wenaus
BNL

RHIC Computing Advisory Committee Meeting
BNL
December 3-4, 1998

Outline

MDC1 in STAR
- Goals
- MDC1 program and results

MDC1 outcomes and coming activity
- Computing decisions

Revisited Computing Requirements
- Status of 3/98 requirements report
- Year one requirements report

Conclusions
- Comments on facilities
- MDC2 objectives

MDC1 Goals

Official throughput goal: processing of 100k events from simulation through the production chain; mix of year 1 and year 2 detector configurations

Beyond the official 100k, a physics driven data set of 300k events

- 100k event goal met; 260k simu, 190k reco total to date
- MDCs a very effective motivator and driver for subsystem software development

Calibration database in place and in use in offline codes

- Was in place, but little used; full deployment for MDC2

Production principally in STAF framework, but ROOT based production exercised as well

- STAF and ROOT both used; ~4% via ROOT to date (manpower limited)

Partial, prototype C++/OO data model in place

- Not met! No manpower. A principal MDC2 goal.

Prototype Objectivity-based DST event store in place

- Met; 50GB of DSTs in Objectivity

Grand Challenge software tested with Objectivity DSTs

- Met; Doug will address

DST based physics analysis

- Low level activity during MDC1; has ramped up post-MDC1

MDC1 Program and Results

Organization

RCF support

Simulations and data sinking

Reconstruction

DST generation, Objectivity based event store

Data mining and physics analysis

Validation and quality assurance

Organization

MDC1 Production Leader

- Yuri Fisyak (BNL)

Simulation and raw data sinking

- Peter Jacobs (LBNL), Pavel Nevski (BNL), Dan Russ (CMU), T3E Team

Production operations

- Lidia Didenko (BNL)

Production chain kumacs, quality control

- Kathy Turner (BNL)

Framework (STAF) development and support

- Victor Perevoztchikov (BNL), Herb Ward (UT Austin)

Grand Challenge data mining

- Doug Olson, David Zimmerman, Ari Shoshani (all LBNL)

Objectivity event store

- Torre Wenaus (BNL)

DST analysis

- A. Saulys, M. Tokarev, C. Ogilvie

And many others!

RCF Support

We found planning and cooperation with RCF and the other experiments before and during MDC1 to be very effective

Thanks to the RCF staff for excellent MDC1 support around the clock and seven days a week!

Particular thanks to...

- HPSS
  - John Riordan
- AFS on CRS cluster
  - Tim Sailer, Alex Lenderman, Edward Nicolescu
- CRS scripts
  - Tom Throwe
- LSF support
  - Razvan Popescu
- All other problems
  - Shigeki Misawa

Simulations and data sinking

Cray T3E at Pittsburgh Supercomputing Center (PSC) became available for STAR last March, in addition to NERSC, for MDC directed simulation

Major effort since then to port simulation software to T3E

- CERNLIB, Geant3, GSTAR, parts of STAF ported in BNL/LBNL/PSC joint project
- Production management and data transport (via network) infrastructure developed; production capability reached in August

90k CPU hrs used at NERSC (exhausted our allocation)

- 50k CPU hrs (only!) allocated in FY99

Production continuing at PSC, working through a ~450k hr allocation through February

- Prospects for continued use of PSC beyond February completely unknown; dependent on DOE

T3E data transferred to RCF via net and migrated to HPSS

- 50GB/day average transfer rate; 100GB peak

Reconstruction

CRS-based reconstruction production chain:

- TPC, SVT, global tracking

Two MDC1 production reconstruction phases

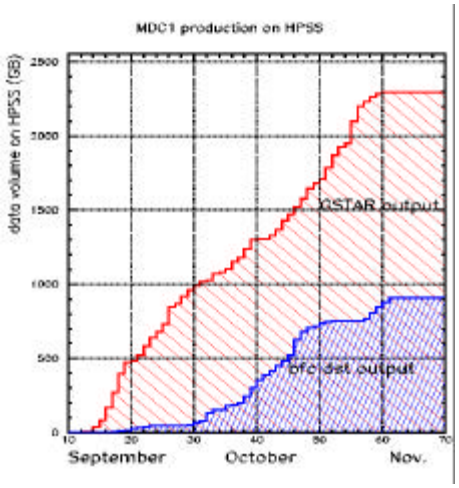
- Phase 1 (early September)
  - Production ramp-up
  - Debugging infrastructure and application codes
  - Low throughput; bounded by STAR code
- Phase 2 (from September 25)
  - New production release fixing phase 1 problems
  - STAR code robust; bounded by HPSS and resources
  - HPSS:
    - size of HPSS disk cash (46GB)
    - number of concurrent ftp processes
  - CPU :
    - from 20 CPUs in early September to
    - 56 CPUs in late October

Production Summary

Total on HPSS: 3.8TB

Geant simulation:
Total: 2.3 TB
NERSC: .6 TB
PSC: 1.6 TB
RCF: ~.1 TB

DSTs:
XDF: 913 GB in 2569 files
Objectivity: 50 GB
ROOT: 8 GB

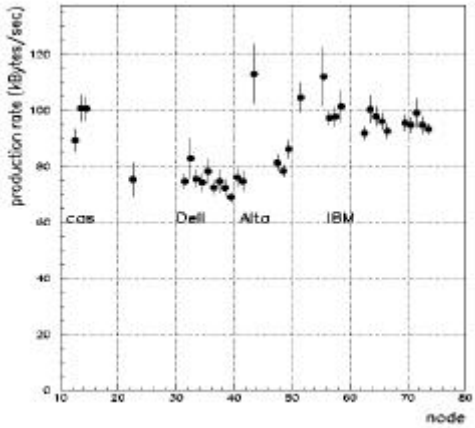


Reconstruction Production Rate

3/98 requirements report:
2.5 kSi95 sec/event
=70 kB/sec

Average rate in MDC1:
75-100 kB/sec

Good agreement!



DST Generation

Production DST is STAF-standard (XDR-based) XDF format

- Objectivity only in secondary role, and presently incompatible with Linux based production

Direct mapping from IDL-defined DST data structures to Objectivity event store objects

- Objectivity DB built using BaBar event store software as basis
- Includes tag database info used to build Grand Challenge query index
- XDF to Objectivity loading performed in post-production Solaris-based step

50GB (disk space limited) Objy DST event store built and used by Grand Challenge to exercise data mining

Data mining and physics analysis

Data mining in MDC1 by Grand Challenge team; great progress, yielding an important tool for STAR and RHIC (Doug's talk)

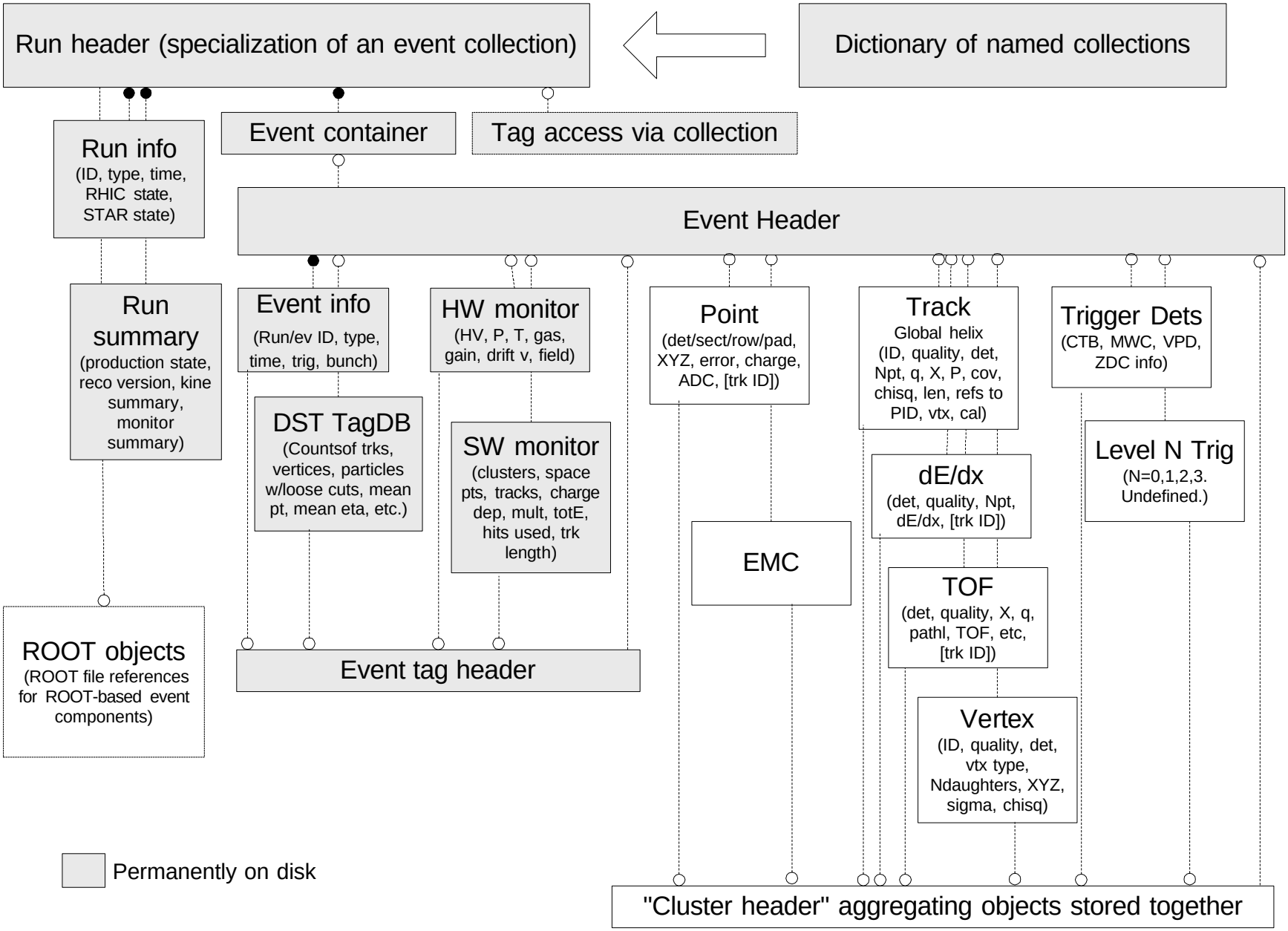
We are preparing for the loading of the full 1TB DST data set into a second generation Objectivity event store

- to be used for physics analysis via the Grand Challenge software
- supporting hybrid non-Objectivity (ROOT) event components
- 'next week' for the last month! Soon...

Little physics analysis activity in MDC1

- Some QA evaluation of the DSTs
- Large MDC1 effort in preparing CAS software by the event by event physics group
  - but they missed MDC1 when it was delayed a month; they could only be here and participate in August
- DST evaluation and physics analysis has ramped up since MDC1 in most physics working group and will play a major role in MDC2

Objectivity-Based Persistent DST Event Model for MDC1



Validation and Quality Assurance

Simulation

- Generation of evaluation plots and standard histograms for all data an integral part of simulation production operations

Reconstruction

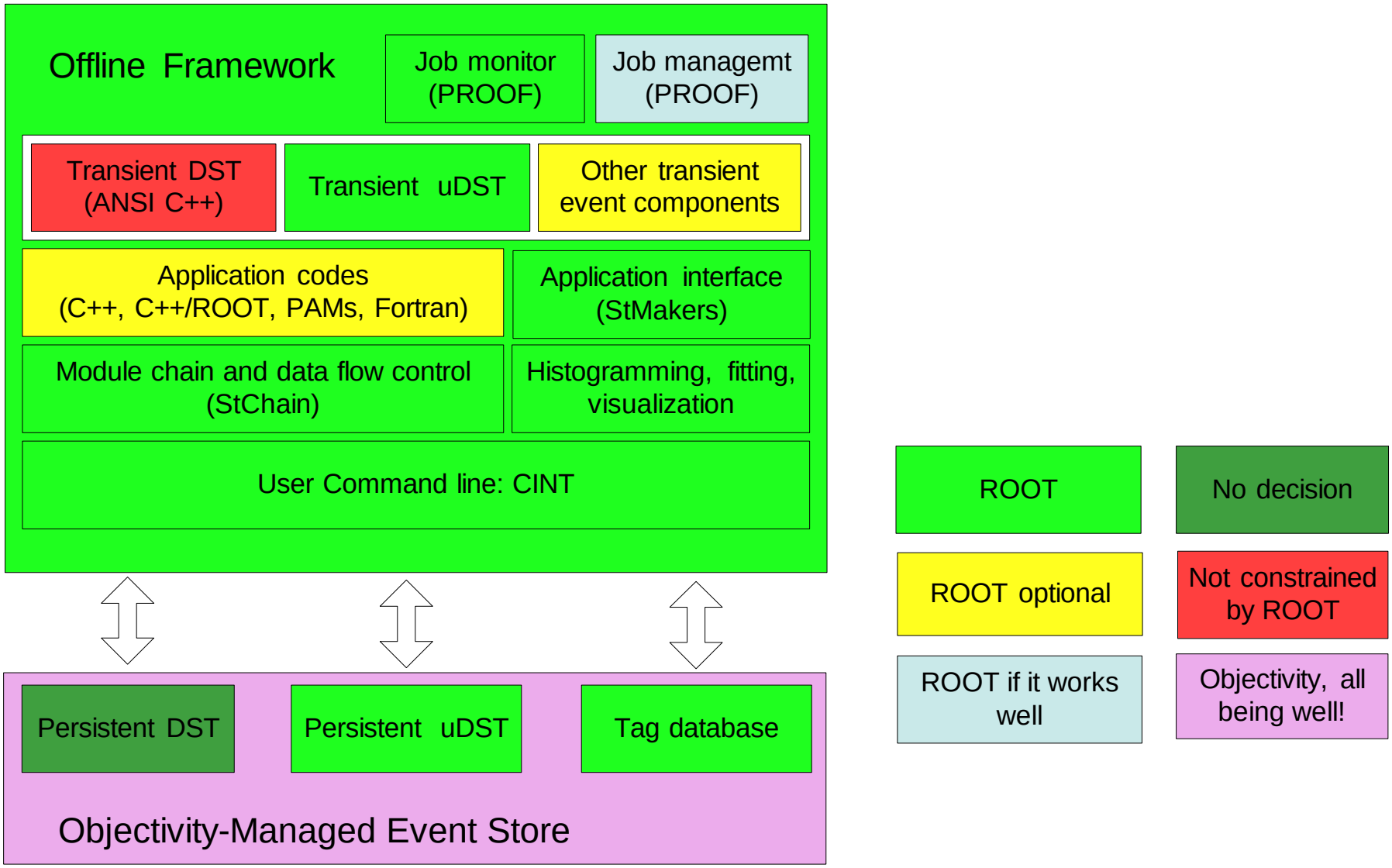
- Leak checking, table integrity checking built into framework and effective in debugging and making the framework leak-tight; no memory growth
- Code acceptance and integration into production chain performed centrally at BNL; log file checks, table row counts
- TPC and other subsystem responsibles involved in assessing and debugging problems
- DST based evaluation of upstream codes, physics analysis play a central role in QA, but limited in MDC1 (lack of manpower)
- Dedicated effort underway for MDC2 to develop framework for validation of reconstruction codes and DSTs against reference histograms

MDC1 Outcomes and Coming Activity

Computing decisions directing development for MDC2 and year 1

- MDC1 followed by intensive series of computing workshops addressing (mainly infrastructure) software design and implementation choices in light of MDC1 and MDC1-directed development work
- Objective: stabilize infrastructure for year 1 by MDC2; direct post-MDC2 attention to physics analysis, reconstruction development
- Many decisions ROOT-related; summarized in ROOT decision diagram
  - Productive visit from Rene Brun
- C++/OO decisions: DST and post-DST physics analysis will be C++ based
- Event Store: Hybrid Objectivity/ROOT event store will be employed; ROOT used for (at least) post-DST micro, nano DSTs; tag database

Proposed ROOT Decisions for MDC2 and Year 1



Revisited Computing Requirements

STAR Offline Computing Requirements Task Force Report

- P. Jacobs et al, March 1998, Star Note 327, 62pp
- Physics motivation, data sets, reconstruction, analysis, simulation
- Addresses fully instrumented detector in nominal year

STAR Offline Computing Requirements for RHIC Year One

- T. Ullrich et al, November 1998, Star Note xxx, 5pp
- MDC1 experience shows no significant deviation from SN327
  - Conclude that numbers remain valid and estimates remain current
- This new report is an addendum addressing year 1
- Detector setup: TPC, central trigger barrel, 10% of EMC, and (probably) FTPC and RICH
- 20MB/sec rate to tape; central event size 16 +- 4 MB
  - No data compression during first year

Year 1 Requirements (2)

RHIC running

- STAR's requirements do not depend on luminosity profile
  - Can saturate on central events at 2% of nominal RHIC luminosity
- Rather, on the actual *duty factor profile*
- 4M central Au-Au at 100 GeV A/beam expected from official estimated machine performance folded with STAR duty factor
  - 25% of nominal year
  - Total raw data volume estimate: 60TB

DST production

- MDC1 experience confirms estimate of 2.5 kSi95/event reconstruction
- Assuming reprocessing factor of 1.5 (very lean assumption for year 1!) yields 15M kSi95 sec
  - 630 Si95 units average for the year, assuming 75% STAR/RCF duty factor
  - Factoring in machine duty factor, can expect bulk of data to arrive late in year

Year 1 Requirements (3)

- To arrive at a reasonable number: simulated a range of RHIC performance profiles
  - Obtained 1.5-2 times average value to achieve year 1 throughput with at most 10% backlog at year end: ~1200 Si95
- DST data volume estimated at 10TB
  - Data reduction factor of 6 rather than 10 (SN327); additional information retained on DST for reconstruction evaluation and debugging, detector performance studies

Data mining and analysis

- Projects able to participate with year 1 data scaled by event count relative to SN327 nominal
  - 12M Si95 sec total CPU needs; 2.3TB microDST volume
  - 480 Si95 units assuming 75% STAR/RCF duty factor
  - Analysis usage is similarly peaked late in the year following QA, reconstruction validation, calibration
    - Need estimated at twice the average, ~1000 Si95 units

Year 1 Requirements (4)

Simulations

- Used to derive acceptance and efficiency corrections, signal background due to physics or instrumental response
  - Largely independent of data sample size; no change relative to nominal year requirements foreseen
  - Data mining and analysis of simulations will peak late in year 1, with the real data
- Simulation CPU assumed to be available offsite
- At RCF, early year CPU not needed yet for reco/analysis can be applied to simulation reconstruction; no additional request for reconstruction of simulated data
- 24TB simulated data volume

Year 1 requirements summarized in report table, p.5

Comments on Facilities

All in all, RCF facilities performed well

- HPSS gave the most trouble, as expected, but it too worked
  - Tape drive bottleneck will be addressed by MDC2

Robust, flexible batch system like LSF well motivated on CAS

- Many users and usage modes

Need for such a system on CRS not resolved

- Limitations encountered in script based system: control in prioritizing jobs, handling of different reco job types and output characteristics, control over data placement and movement, job monitoring
  - Addressable through script extensions? Complementary roles for LSF and in-house scripts?
- Attempt to evaluate LSF in reconstruction production incomplete; limited availability in MDC1
- Would like to properly evaluate LSF (as well as next generation scripts) on CRS in MDC2 as basis for informed assessment of its value (weighing cost also, of course)
- Flexible, efficient, robust application of computing resources important
  - Delivered CPU cycles (≤ available cycles) is the bottom line

MDC2 Objectives

- Full year1 geometry (add RICH, measured B field, ...)
- Geometry,parameters, calibration constants from database
- Use of DAQ form of raw data
- New C++/OO TPC detector response simulation
- Reconstruction production in ROOT
- TPC clustering
- All year 1 detectors on DST (add EMC, FTPC, RICH, trigger)
- ROOT based job monitor (and manager?)
- Much more active CAS physics analysis program
- C++/OO transient data model in place, in use
- Hybrid Objectivity/ROOT event store in use
  - Independent storage of multiple event components
- Quality assurance package for validation against reference results

Conclusions

MDC1 a success from STAR's point of view

- Our important goals were met
- Outcome would have been an optimist's scenario last spring
- MDCs a very effective tool for guiding and motivating effort

Good collaboration with RCF and other experiments

- Demonstrated a functional computing facility in a complex production environment
- All data processing stages (data transfer, sinking, production reconstruction, analysis) successfully exercised concurrently

Good agreement with expectations of our computing requirements document

Clear (but ambitious) path to MDC2

Post-MDC2 objective: a steady ramp to the commissioning run and the real data challenge in a stable software environment